Det egna projektet steg 2

Kommentar

Som dataset valde jag en fil innehållande drygt 3100 matcher mellan 2019 och början av 2024.

Jag ville testa två olika algoritmer för att beräkna matchutgång och valet föll på RandomForest och LogisticRegression i sklearn.

Jag har testat med ett antal olika kombinationer av parametrar från datasetet och utifrån de testerna lagt till de parametrar som gett bäst resultat (finns sannolikt bättre kombinationer, bara testat några stycken).

Sammanfattningsvis kan jag konstatera att båda algoritmer ger liknande resultat, med litet övertag för LogisticRegression. Båda verkar haft betydligt svårare att beräkna oavgjorda resultat än övriga varianter, vilket säkert delvis kan förklaras av att antalet oavgjorda matcher i listan är betydligt färre än vinst/förlust.

För att avläsa ML-resultatet använder jag classification_report och confusion_matrix från sklearn.

Slutligen ville jag se hur viktig varje parameter var vid beräkningen. En fundering jag länge har haft är hur huruvida vem som dömer påverkar matchutgången. Här är dataunderlaget alldeles för litet för att ta reda på det, men på sikt är det något jag ska jobba vidare med. I det här datasetet verkar det spela väldigt lite roll, utan det är som väntat mest xG och xGA som betyder något.

Tillägg 6 mars:
Kom på att jag gjort ett rejält tankefel och använt mig av ett antal parametrar som finns tillgängliga i det historiska datat, men inte kommer finnas tillgängligt före framtida matcher. XG, xGa, poss, sh, dist, sot är alla värden som vi vet först efter matchen. Gjorde en plmatcher_v2.py som bortser från dessa parametrar. Precisionen sjönk som väntat avsevärt, nere på runt 50% nu.

För att lista viktning av varje parameter använde jag feature_importances_ från RandomForest, och matar ut det till en fil som heter importances.txt.